Psychophysical quantification of individual differences in timbre perception

Stephen McAdams & Suzanne Winsberg

IRCAM-CNRS
1 place Igor Stravinsky
F-75004 Paris
smc@ircam.fr

SUMMARY

New multidimensional scaling techniques can be applied to the analysis of dissimilarity judgments on musical timbres in both group and individual data. These techniques make use of objective knowledge we have acquired on the potential physical correlates of the perceptual dimensions that define timbre in a so-called "timbre space". The CONSCAL technique developped by Winsberg & De Soete (1997) constrains the resulting spatial model such that the dimensions correspond to these previously established objective attributes, and such that the order of items along a given perceptual dimension preserves their order along these established physical dimensions. The order-preserving transformation is represented by a monotone spline function and yields what we subsequently interpret as the auditory transform that converts the physical dimension into a perceptual one. A reanalyis of timbre data from the literature demonstrates that this kind of model reveals large differences in the nature of the underlying dimensions as well as in the form of the auditory transforms for different listeners. An analysis of individual data also helps us understand why the higher dimensions in group timbre spaces published in the literature are sometimes difficult to interpret psychophysically.

In A. Schick, M. Meis & C. Reckhardt (Eds.) (2000). *Contributions to Psychological Acoustics: Results of the 8th Oldenburg Symposium on Psychological Acoustics*, Bis, Oldenburg, pp. 165-182.

INTRODUCTION

Timbre is a word used to refer to a collection of auditory attributes that have been approached with many different experimental methods. Some involve deciding a priori what a given attribute is and then proceeding to explore it with unidimensional psychophysical scaling techniques. For example, one might be interested in roughness or sharpness and proceed to evaluate the relative roughness or sharpness of various sounds and then try to link the subjective judgments to physical quantities derived from the sound signals. However this approach presumes on the one hand that listeners know what is meant by the word presented to them, can focus on that attribute and ignore others possessed by the sound, and that they all make the same link between the word and a specific aspect of their perception. This approach also presumes that psychoacousticians are clever enough to imagine what all the attributes might be ahead of time in order to specifically and directly test them in such a way. Both of these assumptions do not always hold true. It is sometimes difficult to get listeners to understand what aspect of perception corresponds to a given word. And there may be perceptual attributes that are part of complex sounds that scientists have not yet thought of investigating systematically.

Multidimensional scaling (MDS) of dissimilarity judgments provides an exploratory data analysis tool for discovering what aspects of sound listeners use to compare sounds without having any a prioris concerning what these aspects might be (Plomp, 1970; Grey, 1977; Krumhansl, 1989; Iverson & Krumhansl, 1993; McAdams, Winsberg, Donnadieu, De Soete & Krimphoff, 1995). And when combined with acoustic analysis and psychoacoustic modeling, this approach can even give rise to psychophysical quantification of the perception dimensions that have been discovered (Grey & Gordon, 1978; Iverson & Krumhansl, 1993; Krimphoff, McAdams & Winsberg, 1994). We will briefly present the MDS approach as applied to data for groups of listeners using the CLASCAL technique (Winsberg & De Soete, 1993) that presumes nothing about the physical structure of the sounds being judged. From the acoustic analyses of the dimensions thus revealed we will then present a new approach, CONSCAL (Winsberg & De Soete, 1997), in which the MDS analysis is constrained by physical parameters that are known to be used by listeners for a given set of sounds. We will show that this approach is particularly useful in describing individual psychophysical functions on multiple perceptual attributes.

MULTIDIMENSIONAL SCALING WITH CLASCAL

In our experiments on the perception of musical timbre (McAdams et al., 1995), the aim has been to determine the structure of the multidimensional perceptual representation of timbre, or what has come to be called "timbre space", for individual notes played by musical instruments and then to attempt to define the acoustic and psychoacoustic factors that underly this representation. The combination of a quantitative model of perceptual relations among timbres and the psychophysical

explanation of the parameters of the model is an important step in gaining predictive control of timbre in several domains such as sound analysis and synthesis and intelligent search in sound databases. Of course, such representations are only useful to the extent that they are: 1) generalizable beyond the set of sounds actually studied, 2) robust with respect to changes in musical context, and 3) generalizable to other kinds of listening tasks than those used to construct the model. To the degree that a representation has these properties, it may be considered as a genuine model of musical timbre, the main feature of a good model being predictive power.

The development of techniques for multidimensional scaling (MDS) of proximity data in the 1950s and 1960s have provided a tool for exploring complex sensory representations (see McAdams et al. 1995, for a review). These techniques have several advantages as well as a few limitations. The primary advantage is that from a relatively simple task—judging the degree of similarity or dissimilarity between all pairs of stimuli from a fixed set—an ordered structure is obtained that can often lend itself to psychophysical quantification. Applied for the first time to musical timbre by Plomp (1970) and subsequently by Wessel (1973) and Miller and Carterette (1975), this kind of analysis searches for structure in the perceptual data without obliging the experimenter to make any a priori assumptions about the nature of that structure. Often, we are interested in discovering the perceptual structure of a set of complex sound events, the nature of which we do not know in advance. These techniques are quite useful for this kind of exploratory data analysis, although they can also be used for more confirmatory analyses, once one has a more clear idea of the relations among acoustic and perceptual parameters.

The basic principle of MDS is the following. A set of stimuli (for example sounds equalized in pitch, loudness, duration, and spatial position) is presented to a group of listeners in all possible pairs. The listeners are asked to rate the degree of dissimilarity between each pair of timbres on a numerical scale or with a continuous slider. This scale gives high similarity at one end and high dissimilarity at the other. The basic assumption is that there exists a mental representation of each timbre that has certain prominent components and the number or slider position reflects a comparison based on these components. Furthermore, this representation is assumed to be relatively similar across listeners (perhaps with some variations that will be discussed below). So the structure in the data should somehow reflect the perceptual structure. The data set for each listener can be arranged in the form of a matrix, each cell corresponding to a pair of timbres. The matrix or set of matrices from different subjects or conditions are analyzed in an MDS program, the main task of which is to fit a distance model to the dissimilarity data so that a monotonic or linear relation exists between the two, i.e. the greater the dissimilarity, the greater the distance.

Goodness-of-fit statistics are used to determine the number of dimensions to retain, and also, in the case of a weighted model in which different subjects or classes of subjects weight the dimensions differently, the psychologically meaningful dimensions to interpret. The various techniques differ in terms of 1) the spatial models that are evaluated, 2) the loss function used to measure the goodness-of-fit of the model to the data, 3) the numerical algorithm used to find the parameters of the model. We prefer maximum likelihood methods allowing model selection using log likelihood-based information criteria (BIC) and Monte Carlo tests.

We often use the CLASCAL program for MDS analysis (Winsberg & De Soete, 1993). This program uses a maximum likelihood procedure for fitting an extended Euclidian distance model to dissimilarity judgments made by a set of listeners on all pairs of sounds from a predetermined set. The principle behind the analysis is that listeners use a small number of perceptual dimensions or features associated with the sounds to judge how similar or dissimilar they are. It also presumes that this set of perceptual dimensions and features is the same for all listeners, with the possibility that different classes of listeners will weight the various dimensions and set of features on individual sounds in different ways. Part of the output of the algorithm is a set of coordinates in a Euclidean space. The model thus presumes that the timbres share all the perceptual dimensions. However, in some cases the stimuli, sounds in our case, may have characteristics that no other sounds in the set have (like the rapid damping of a harpsichord sound or the weak even-numbered harmonics in a clarinet sound). These sounds have "specificities" that make them dissimilar to all the other timbres, but such features cannot be accounted for by the shared dimensions along which vary all the timbres of the tested set in a continuous fashion. There are two possible sources for such specificities. Either a given specificity represents an additional dimension along which only one timbre varies, or it represents one or more features not present in the rest of the sounds. So the Euclidean distance model is extended to include specificities on individual timbres in addition to their common dimensions.

Finally, we consider that different subjects may weight the different dimensions and specificities according to their perceptual salience and that subjects form "latent classes" that can be determined on the basis of their data. The classes are "latent" in the sense that they are not predetermined but are derived from the data. This latent-class approach was implemented in the CLASCAL program by Winsberg and De Soete (1993). The appropriate number of latent classes is determined and statistical tests are also performed to estimate the probability that each subject belongs to each class. In general subjects are assigned to a single class, although class belongingness can be ambiguous for some subjects. The combination of the extended Euclidean model and the latent-class approach has resulted in an extension of the CLASCAL model. This distance model has both specificities and class weights; the weights are applied to each dimension and to the set of specificities taken collectively. In this model, the distance between stimuli *i* and *j*, d_{ij}, is given by:

$$d_{ij} = \left[\sum_{1}^{K} w_{kc}\left(x_{ik} - x_{jk}\right)^2 + v_c\left(s_i + s_j\right) \right]^{\frac{1}{2}}, \qquad (1)$$

where x_{ik} is the coordinate of timbre *i* on dimension *k*, s_i is its specificity, w_{kc} is the weight on dimension *k* for class *c* and v_c is its weight on the set of specificities.

This model was used by McAdams et al. (1995) to study a set of 18 musical instruments synthesized with frequency modulation algorithms developed by Wessel, Bristow and Settel (1987) on a Yamaha synthesizer. These instruments were intended either to imitate conventional orchestral instruments or to constitute chimeric hybrids

between them (e.g., the *vibrone* is a hybrid between vibraphone and trombone). All pairs of sounds were presented to 84 listeners who judged their relative dissimilarity on a numerical scale from 1 (very similar) to 9 (very dissimilar). In reanalyzing the data from the 24 professional musicians among those subjects, the CLASCAL analysis revealed a three-dimensional space without specificities and two latent subject classes. Figure 1 presents this timbre space. Note that while the timbres are distributed in a relatively homogeneous manner along Dimensions 2 and 3, they form two large clusters along Dimension 1.

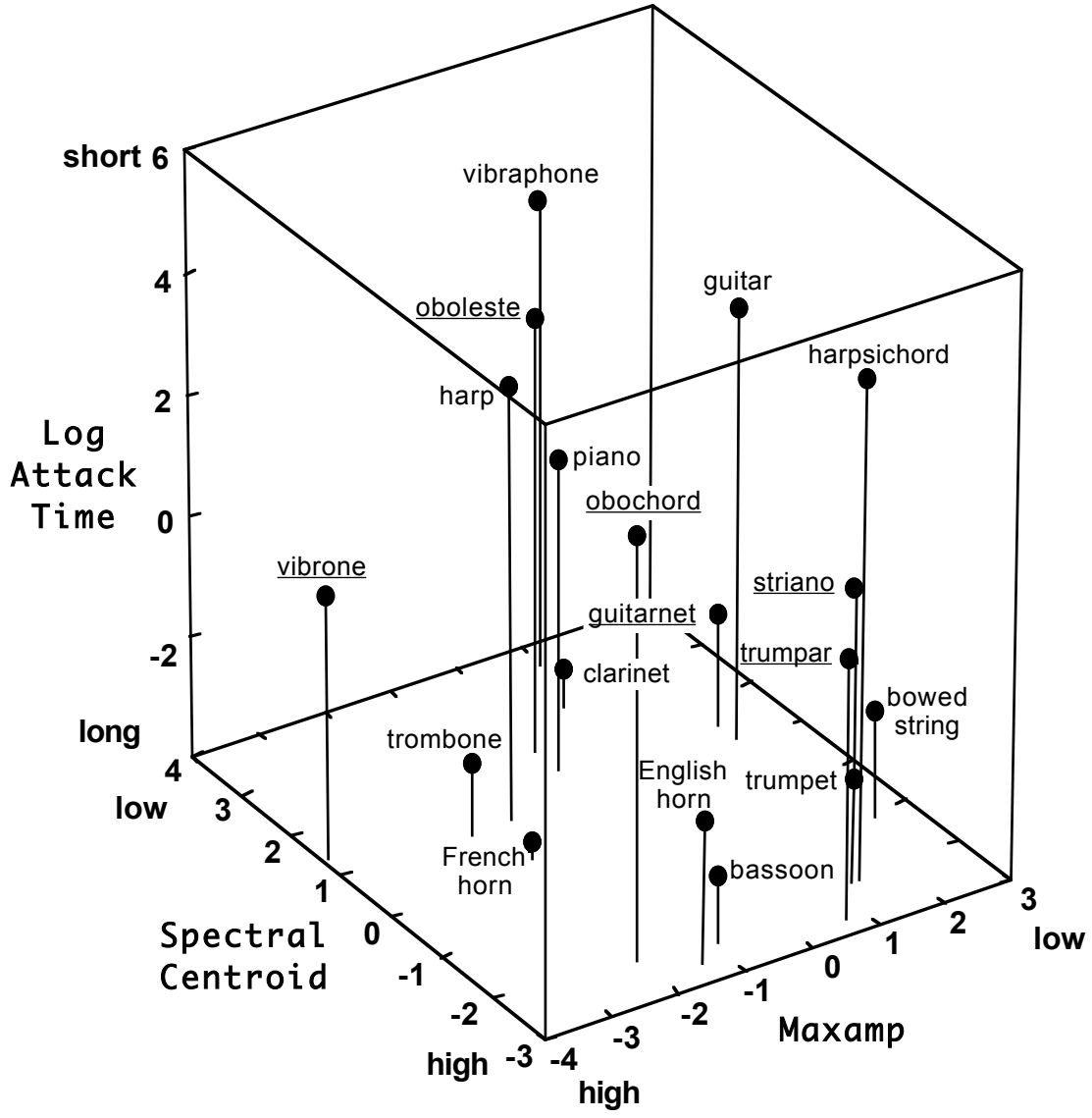

FIGURE 1. A three-dimensional timbre space found from a CLASCAL analysis on dissimilarity data from 24 professional musicians that formed two latent classes. Underlined instrument names represent hybrids (oboleste = oboe+celeste, obochord = oboe+harpsichord, vibrone = vibraphone+trombone, striano = *bowed* string+piano guitarnet = guitar+clarinet, trumpar = trumpet+guitar). The corresponding acoustical correlates of each perceptual dimension are indicated in parentheses.

Individual CLASCAL analyses on each listener's data were also performed. We examined the best two models selected by the BIC statistic. For the best model, 13 of the 24 subjects had one-dimensional solutions (11 with specificities), seven had two-dimensional models without specificities and four had three-dimensional models without specificities. It is very interesting to note that the individual dimensionalities are generally much lower than the group dimensionality.

ACOUSTICAL CORRELATES OF PERCEPTUAL DIMENSIONS

Our approach to determining the acoustic correlates of timbre space focused initially on the spaces of Krumhansl (1989) and McAdams et al. (1995) using the FM timbres, but has expanded more recently to include several other spaces, using analyzed/resynthesized or recorded sounds, that have been published in the literature or are currently submitted for publication (McAdams and Winsberg, in preparation; McAdams, Susini, Krimphoff, Misdariis & Smith, in preparation). We tend to use an empirical loop consisting of listening to the sounds in front of a visual representation of the timbre space and to try to get an auditory sense of what changes systematically as one plays a timbre trajectory across a given dimension. The initial impression then leads to the development of signal-processing algorithms, usually based on a time-frequency representation derived from a short-term Fourier analysis (phase vocoder and the like). We have used both the *Additive* environment developed at IRCAM (Depalle, García & Rodet, 1993) and Beauchamp's (1993) *Sndan* environment. The goal is to find a parameter that varies in a linear relation with the coordinates of the timbres along a given dimension in the timbre space. So we try various algorithms that provide a single parameter per sound and then either reject them or progressively refine them until the correlations are as high as possible. This approach was first applied by Krimphoff et al. (1994) to Krumhansl's (1989) space. The main four correlates are specified in Equations 2-5 (LAT=Log Attack Time, SC=Spectral Centroid, SS=Spectral Smoothness, and SF=Spectral Flux). Attack time is the time it takes to progress from a threshold energy level to the maximum in the rms amplitude envelope. Spectral centroid is the center of gravity of the long-term amplitude spectrum. Spectral smoothness is related to the degree of amplitude difference between adjacent partials in the spectrum computed over the duration of the tone. A trumpet often has a smooth spectrum and a clarinet a jagged one, so the former would have a low value of SS and the latter a higher one. Spectral flux is a measure of the degree of variation of the spectrum over time.

$$LAT = \log_{10}(t_{max} - t_{threshold}) \quad (2)$$

$$SC = \frac{1}{T}\int_0^T B(t)dt \quad \text{with} \quad B(t) = \left[\frac{\sum_{k=1}^{N} kA_k(t)}{\sum_{k=1}^{N} A_k(t)} \right] \quad \text{for a given analysis window} \tag{3}$$

$$SS = \sum_{k=1}^{N} \left| 20\log(A_k) - \frac{20\log(A_{k-1}) + 20\log(A_k) + 20\log(A_{k+1})}{3} \right| \tag{4}$$

$$SF = \frac{1}{M}\sum_{p=1}^{M} \left| r_{p,p-1} \right| \quad \text{with} \quad M = \frac{T}{\Delta t} \quad \text{and} \quad \Delta t = 16ms \tag{5}$$

where t_{max} is the instant in time at which the rms amplitude envelope attains its maximum, $t_{threshold}$ is the time at which the envelope exceeds a threshold value ($0.02*t_{max}$ in our case), T is the total duration of the sound, t is the begin time of the sliding short-term Fourier analysis window, A_k is the amplitude of partial k, N is the total number of partials, $r_{p,p-1}$ is the Pearson product-moment correlation coefficient between the amplitude spectra at times t_p and t_{p-1}.

For this particular timbre space based on group data, we found very high correlations with log attack time (LAT, r=0.94, Dim1) and spectral centroid (SC, r=.90, Dim2) for two dimensions and a relatively high one with the maximum instantaneous amplitude attained by the energy envelope of the signal (maxamp, r=.73, Dim3). Lower correlations were found with other factors: Dim1 was well correlated with the effective duration measured at –3dB from the maximal level in the rms amplitude envelope (r=.81), Dim2 was weakly correlated with spectral smoothness (r=.46), and Dim3 was weakly correlated with spectral flux (r-.43). In Krumhansl's (1989) space, one dimension was temporal (LAT) and two were spectral in nature (SC and SS). High correlations were also found for LAT with Dim1 and SC with Dim2 in the McAdams et al. (1995) space with all 84 listeners. However, Dim3 in this latter space was spectro-temporal in nature and was correlated (somewhat more weakly) with SF.

For the individual timbre spaces, LAT explained the first dimension for 23 of the 24 listeners. SC explained the first dimension for one listener, the second dimension for seven of the 11 listeners with two- or three-dimensional spaces and the third for another. Maxamp explained the second dimension for one listener and the third dimension for three of the four listeners having three dimensions. As we can see, there is a preponderance of LAT and SC in the physical parameters that make evident the source of these dimensions in the group space. The lower correlation with maxamp for the third dimension of the group space is explained by its importance for a small number of listeners. However, the fact that it shows up in the group space is perhaps due to the

fact that it predominates the third dimension among listeners having this many dimensions.

CONSTRAINED MULTIDIMENSIONAL SCALING WITH CONSCAL

It is at times difficult to determine the appropriate dimensionality based on goodness-of-fit statistics. The unweighted distance model is rotationally invariant, so if the unweighted model has been used, it is often difficult to find a rotation such that all dimensions are interpretable. Even when the weighted model is used, removing rotational invariance, it is sometimes difficult to interpret all of the recovered "psychological" dimensions. Moreover, this problem may occur in situations where a small number of physical parameters can be used to describe the objects. In such a case it may be more fruitful to use the information at hand and constrain the dimensions of the distance model to be monotone transformations of these physical dimensions. This is what the CONSCAL program (Winsberg & De Soete, 1997) does.

CONSCAL constrains the resulting spatial model such that the order of items along a given perceptual dimension preserves their order along a previously established physical dimension. The fit between perceptual and physical dimensions is achieved with monotone spline functions and yields what may be interpreted as the auditory transform of the physical dimension needed to obtain the perceptual one. The distance model in CONSCAL has the following form for the case of an identity metric in which the dimensions are orthogonal:

$$d_{ij} = \left[\left(\mathbf{f}_i - \mathbf{f}_j \right)^T \; I \left(\mathbf{f}_i - \mathbf{f}_j \right) \right]^{1/2} = \left[\sum_{k=1}^{K} \left(\mathbf{f}_i - \mathbf{f}_j \right)^2 \right]^{1/2}, \qquad (6)$$

There are *K* dimensions and the physical predictor variable *k* is denoted by superscript *(k)*. I is the *K*×*K* identity matrix. $\mathbf{f}_i$ is the set of perceptual coordinates for timbre *i*, represented as the vector of monotone transformations for timbre *i*, the k^{th} component being $f_i^{(k)}\left(x_i^{(k)}\right)$, where $f^{(k)}(\cdot)$ is the spline monotone transformation for dimension *k* and $x_j^{(k)}$ is the physical coordinate of object *i* on dimension *k*. The transformation function for each dimension is defined to be zero at the smallest physical value. A more complex model exists for partially correlated dimensions in which the identity matrix is replaced by a symmetric matrix describing the relative degree of rotation of each axis with respect to each other axis.

A spline function is a piecewise polynomial joined at a finite number of junction points defined over the range of values under consideration. The order of the splines is the maximal degree of the polynomials plus one. In addition to the maximal degree of the splines, the number and location of a strictly increasing number of junction points must be specified in advance, as well as the number of continuous derivatives including the the zeroth derivative (the function), which exist at each junction point. In the

important special case where the spline has maximal continuity equal to the order of the splines at each junction point, the number of parameters required for each dimension is the order plus the number of interior junction points. The number of degrees of freedom in this model is equal to the sum of the number of parameters per dimension across all dimensions. Note that this model is extremely parsimonious compared to classical MDS models since one can add a lot of stimuli and subjects without increasing the number of model parameters, provided that the number of dimensions remains the same and the transformation remains as smooth.

We applied this approach to the group data for the 24 professional musicians comparing the timbre set presented in Figure 1. We tested for the parameters LAT and SCG for dimensions 1 and 2 and tried various physical parameters for dimension 3 (SS, SF, and maxamp). Using Monte Carlo tests, this model was then compared to the CLASCAL model with specificities and latent classes. The CONSCAL model was rejected in favor of the CLASCAL model in all cases. Given that the individual analyses showed differences in dimensionality and in the underlying physical nature of the dimensions across listeners, we selected a subset of nine listeners that had only two dimensions in their individual analyses. Further, these two dimensions always correlated best with LAT and SC. CLASCAL still modeled the data better than CONSCAL. This latter result suggests large differences in the psychophysical functions relating the physical variables to the perceptual dimensions for individual subjects.

We therefore performed the CONSCAL/CLASCAL comparison on the data for individual listeners. For eight of the nine listeners, the CONSCAL model fit the data better than the CLASCAL model, and for the ninth listener the two models were equivalent. This result demonstrates clearly that the CONSCAL approach can be quite useful in modeling the perception of complex sounds for individual data. But why does the group analysis fail? The answer is coherent with the hypothesis that led us to examine the individual analyses and can be gleaned from inspection of Figure 2.

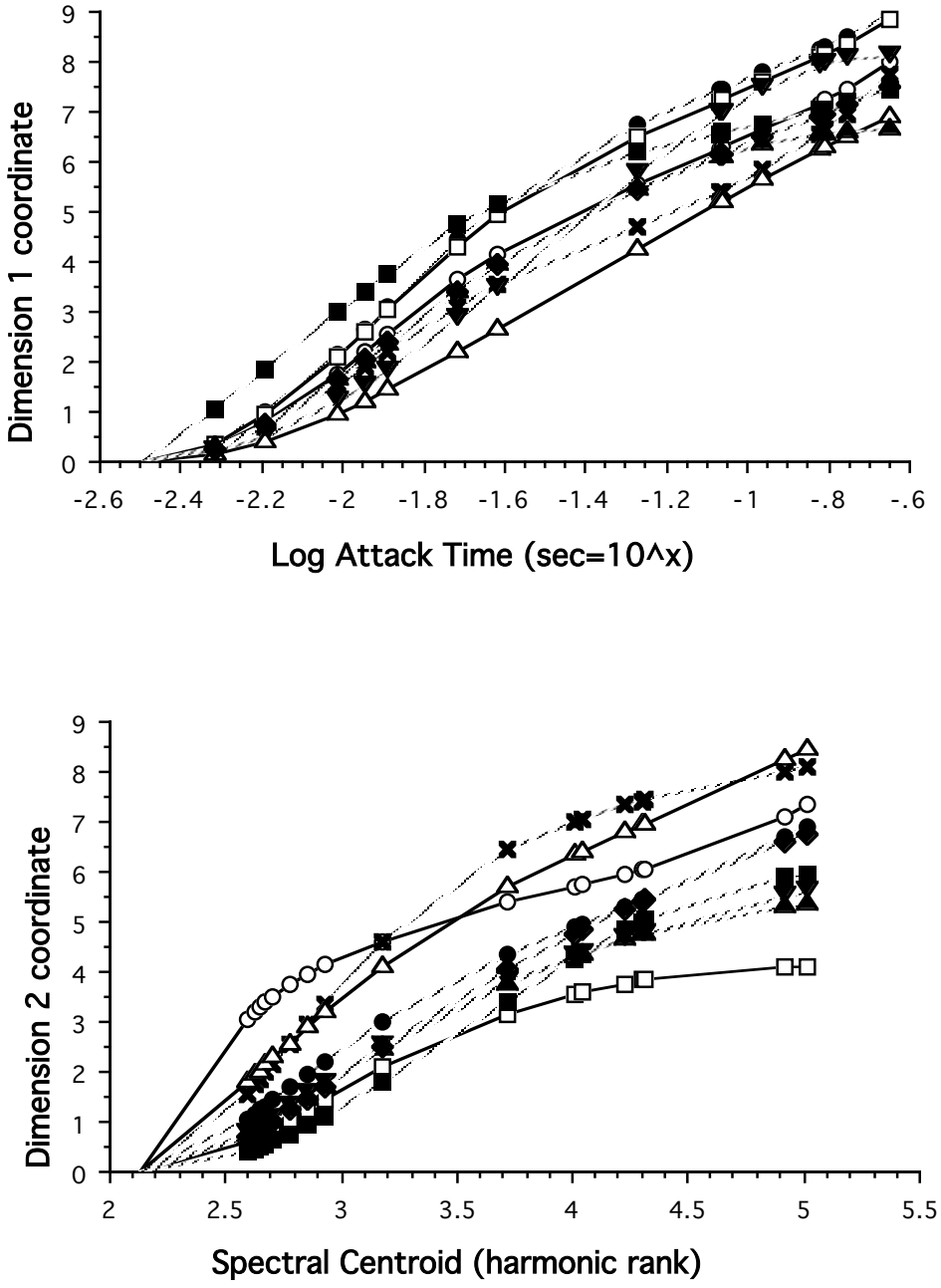

FIGURE 2. Individual psychophysical functions derived with the program CONSCAL for nine musician listeners having two-dimensional perceptual spaces. The upper panel shows the functions for log attack time and the lower one for spectral centroid. Each graph represents the coordinate on the perceptual dimension as a function of the physical value for each of the 18 synthetic timbres. The curves for three listeners discussed in the text are plotted with solid lines and open symbols.

This figure presents the spline functions used to fit the dissimilarity judgments to the physical parameters for each subject. Note that not only is the global weight

attached to each dimension different (as would be estimated for individual subjects by INDSCAL or classes of subjects by CLASCAL), but that the forms of the psychophysical functions are different. To illustrate this point, the functions for three subjects have been highlighted in the figure. Listener L1 (open triangles) has the lowest values for attack time and the function is nearly linear. L1 has the second highest function for spectral centroid also with a nearly linear function. Listener L2 (open squares) has fairly high values for attack time with a slightly compressive function at higher values of this physical variable, while also having very low values for spectral centroid with a strongly compressive function. Finally, listener L3 (open circles) has intermediate values for LAT with a nearly linear function and high values for SC with strong compression at low physical values and a rise at higher values. Thus the forms of these psychophysical functions are very different across individuals, perhaps indicating differences in either judgment strategy or even in perceptual sensitivity to or sensory representation of these physical parameters. At a more global level, this analysis approach also allows us to demonstrate differences in the degree of variability across listeners for a given physical variable. Note that the variation across functions is much smaller for attack time than for spectral centroid.

CONCLUSIONS

The CONSCAL approach to multidimensional psychophysical scaling has demonstrated that previous knowledge of physical parameters can allow the determination of auditory transforms within a multiparameter context. However, this approach does not work as well as the CLASCAL model on group data. The latter approach may work better on group data since it includes specificities and latent class weights, but also because the fitting of spline transformations of physical values to model the perceptual ones is inherently noisy on group data due to individual differences in auditory transforms of physical parameters. When analyzing individual data, to the contrary, good fits are found and the psychophysical functions are well estimated.

ACKNOWLEDGEMENTS

This work has benefitted from collaboration with several colleagues, particularly concerning data collection and the determination of acoustic correlates. We thank Sophie Donnadieu, Jochen Krimphoff, Nicolas Misdariis, Bennett Smith, and Patrick Susini for their helpful input.

REFERENCES

Beauchamp, J. W. (1993,). Unix workstation software for analysis, graphics, modifications, and synthesis of musical sounds. Paper presented at the 94th Convention of the Audio Engineering Society, Berlin.

Depalle, P., García, G., & Rodet, X. (1993,). Tracking of partials for additive sound synthesis using hidden Markov models. Paper presented at the ICASSP.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. Journal of the Acoustical Society of America, 61, 1270-1277.

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. Journal of the Acoustical Society of America, 63, 1493-1500.

Iverson, P., & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. Journal of the Acoustical Society of America, 94, 2595-2603.

Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caractérisation du timbre des sons complexes. II: Analyses acoustiques et quantification psychophysique. Journal de Physique, 4(C5), 625-628.

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In S. Nielzén & O. Olsson (Eds.), Structure and Perception of Electroacoustic Sound and Music, (pp. 43-53). Amsterdam: Excerpta Medica.

McAdams, S., Susini, P., Krimphoff, J., Misdariis, N., & Smith, B. K. (in preparation). A meta-analysis of timbre space. II: Acoustic correlates of common perceptual dimensions. .

McAdams, S., & Winsberg. (in preparation). A meta-analysis of timbre space. I: Multidimensional scaling of group data with common dimensions, specificities and latent subject classes. .

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. Psychological Research, 58, 177-192.

Miller, J. R., & Carterette, E. C. (1975). Perceptual space for musical structures. Journal of the Acoustical Society of America, 58, 711-720.

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In R. Plomp & G. F. Smoorenburg (Eds.), Frequency Analysis and Periodicity Detection in Hearing, (pp. 397-414). Leiden: Sijthoff.

Wessel, D. L. (1973). Psychoacoustics and music: A report from Michigan State University. PACE: Bulletin of the Computer Arts Society, 30, 1-2.

Wessel, D. L., Bristow, D., & Settel, Z. (1987,). Control of phrasing and articulation in synthesis. Paper presented at the 1987 International Computer Music Conference.

Winsberg, S., & De Soete, G. (1993). A latent class approach to fitting the weighted euclidean model. CLASCAL. Psychometrika, 58, 315-330.

Winsberg, S., & De Soete, G. (1997). Multidimensional scaling with constrained dimensions: CONSCAL. British Journal of Mathematical and Statistical Psychology, 50, 55-72.